NOTAS

USOS Y ABUSOS DE LAS CORRELACIONES EN BIOLOGIA

ARTURO I. KEHR

Centro de Ecología Aplicada del Litoral, C.C. 291, (3400) Corrientes, Argentina

Los coeficientes de correlación (r) son muy utilizados en las investigaciones biológicas, y el uso de los mismos se ha generalizado tanto en las investigaciones de índole ecológicas como en las de sistemática. Sin embargo, no siempre son tenidas en cuenta las restricciones de su uso o aquellos requisitos necesarios para una válida interpretación de los mismos. Mi intención en esta Nota es tratar de desarrollar brevemente algunas de las restricciones para el uso de los coeficientes de correlación, como así también en aquellos casos en que fuera posible, la "solución" estadística para el correcto empleo de los mismos. Considero que tener conocimiento sobre esta temática seguramente redundará en la calidad de las interpretaciones que se hagan de los datos, como así también en el diseño de las investigaciones que se quieran realizar.

1) ***Las correlaciones** significativas estadísticamente expresan la existencia de una relación lineal entre dos variables y no implican causalidad.*

En la regresión lineal simple se interpreta la dependencia lineal de una variable (Y) con relacion a otra considerada independiente (X). Sin embargo, la interpretación que podemos hacer del coeficiente de correlación (r) (también llamado coeficiente de correlación simple o coeficiente de correlacion producto-momento de Pearson) es que dicho valor también representa la relación lineal entre dos variables, pero asumiendo que ninguna es dependiente de la otra. Por lo tanto, el coeficiente de correlacion es un valor que representa la intensidad de asociación entre dos variables y no una medida del cambio de una variable con respecto a la otra, lo cual implicaría causalidad. Una correlación entre dos variables (A y B) puede ser producida porque: A influye sobre B, B influye sobre A, o ambas variables están relacionadas a una tercera. Los problemas generalmente comienzan con el último caso. Por ejemplo, existe una correlación positiva entre la edad de un sapo con la densidad de la población mundial. Sin embargo, esto no significa que una variable cause un efecto sobre la otra, sino que ambas están relacionadas a una tercera que es el tiempo.

2) La imposibilidad de obtener un valor promedio de los coeficientes de correlación.

Los coeficientes de correlacion no se comportan como los números ordinarios y por lo tanto no obedecen a las reglas básicas de la aritmética. Resulta incorrecto obtener el promedio de varios coeficientes de correlación si lo hacemos de la manera tradicional (Martin y Bateson, 1993). Para hallar la media aritmética de varios coeficientes de correlación debemos primeramente convertir a cada uno de ellos al valor correspondiente después de la aplicación de la transformación z de Fisher. Dicha fórmula es:

$$z = 1/2 \ln[(1 + r) / (1 - r)]$$

donde ln es igual al logaritmo natural y r es el coeficiente de correlación. Con los valores z calculados es posible obtener la media aritmética. Finalmente, la media aritmética de los valores z debe ser reconvertida en un valor r, a través de la transformación inversa de dicho valor. Tablas para la transformación de r en z y viceversa pueden ser obtenidas en varios libros de estadística básica (por ej., en Snedecor y Cochran, 1980; Zar, 1984).

Ejemplo: Se desea obtener un índice de correlación medio que represente la relación entre las variables ancho máximo de la cabeza : longitud del cuerpo en individuos de una especie de anfibios. Para esto han sido obtenidos varios coeficientes de correlación de individuos pertenecientes a distintas poblaciones.

Coef. de correlación obtenidos:

Población	A	0,89 (n= 25 indiv.)
Población	B	0,92 (n= 15 indiv.)
Población	C	0,93 (n= 20 indiv.)
Población	D	0,90 (n= 17 indiv.)
Poblacion	E	0,93 (n= 22 indiv.)

Después de las transformaciones (z), los valores respectivos obtenidos (considerando solamente dos decimales) fueron: Población A:

1,42; B: 1,58; C: 1,65; D: 1,47 y E: 1,65. La media aritmética obtenida es: $z = 1{,}55$. Finalmente, para calcular la media aritmética de los valores de r, debemos hacer:

$$r = 1 / 1{,}55$$

$$r = 0{,}64$$

La metodología explicada se utiliza con el coeficiente de correlación el cual es un método paramétrico. La mejor manera de obtener la media aritmética de una serie de índices de correlación obtenidos a partir del índice de Spearman (método noparamétrico) es calculando la mediana de dicha serie de valores.

3) ***Los coeficientes de correlación no pueden ser directamente comparados.***

Dos coeficientes de correlación no pueden ser comparados de la misma manera como lo hacemos con datos de peso, longitudes, etc. (Martin y Bateson, 1993). Por ejemplo, un coeficiente de correlación de 0,6 no representa una asociación entre variables dos veces superior que un coeficiente de 0,3. Para comparar una serie de coeficientes de correlación, la mejor manera es utilizar el cuadrado de dichos coeficientes, también llamado coeficiente de determinación. El coeficiente de determinación (r^2) es un coeficiente que representa la proporción de la variación de una variable producida por la variación de la otra variable. De este modo, un coeficiente de correlación de 0,6 ($r^2 = 0{,}36$) significa que el 36% de la variación en una variable es producida por la variación de la otra. Un coeficiente de correlación de 0,3 ($r^2 = 0{,}09$) significa que solamente un 9% de la variación observada en una variable es producida por la variación de la otra. Por lo tanto, y volviendo a la comparación entre distintos coeficientes de correlación, podemos decir que una correlación de 0,6 es cuatro veces superior a una correlación de 0,3.

Lo anteriormente explicado tendrá una real validez siempre que los dos coeficientes comparados sean significativos ($P < 0{,}05$), es decir que ambos coeficientes fueran estadísticamente distintos de O Para poder establecer si existe una correlación entre dos variables, debemos comprobar las dos hipótesis posibles: H_0: $r = 0$ (no existe correlación); HA: $r \neq 0$ (existe una correlación significativa). Esto se puede conocer, utilizando un test de Student, para lo cual es necesario calcular el error standard de r y los grados de libertad (g.l. = n - 2, donde n = número de pares de datos comparados). De este modo, al comparar dos coeficientes de correlación, ambos significativos, estaremos seguro que la asociación observada entre ambas variables, para cada uno de los coeficientes, no se debe solamente a efectos del azar o chance (Rohlf y Sokal, 1969; Zar, 1984).

4) ***Una correlación entre dos variables es válida si aquellas fueron obtenidas de poblaciones con idénticas varianzas y normalmente distribuidas.***

Cuando se interpreta una correlación generalmente asumimos que las variables poseen iguales varianzas o la diferencia entre ellas es pequeña. Sin embargo, frecuentemente esta premisa no se cumple. Esto sucede cuando en un eje de coordenadas es visualizado, por ejemplo, un incremento de la variabilidad del eje Y cuando se incrementa la magnitud del eje X. También en otros casos es frecuente observar, una gran variabilidad entre los datos ubicados en la parte central de la distribución de las variables, mostrando una fuerte asociación solamente en los extremos de las mismas. Como es fácil imaginar, los coeficientes de correlación obtenidos en estos casos son de poca validez. Una manera de comprobar si las varianzas de dos variables son homogéneas (iguales) o heterogéneas, es realizando una prueba de F entre ambas varianzas o el test de Bartlett, de acuerdo a los pasos propuestos por Sokal y Rohlf (1981). Lamentablemente, si bien ambos test son usados frecuentemente, estos resultan ser muy sensibles a la falta de normalidad en los datos de las variables. Por lo tanto, el uso de los mismos es recomendado solamente en aquellos casos cuando se conoce que las variables se distribuyen en forma normal. Un test alternativo para comprobar homogeneidad en las varianzas es el test de Levene (1960), el cual se caracteriza por ser muy robusto. El mismo se basa en transformar a los datos originales en desviaciones absolutas con respecto a la media, para luego aplicar el test de Student, con el motivo de observar diferencias significativas entre las medias aritméticas de las desviaciones, de las dos muestras consideradas. Sin embargo,

el mismo test resulta a su vez más robusto si utilizamos las desviaciones absolutas con respecto a las medianas de cada variable (Schultz, 1983). Si las varianzas fueran heterogéneas, una posibilidad para la solución de este problema es realizar una transformación logarítmica (decimal o natural) de una o ambas variables.

Otra asunción de la cual se parte para una válida interpretación del coeficiente de correlación, es considerar que ambas variables se distribuyen normalmente. En las regresiones uno asume que por cada valor de X (variable independiente) los valores de Y (variable dependiente) han sido tomados al azar de una población normal. A su vez, en las correlaciones, no solamente es asumido que ocurra esto último sino que, además, los valores de X por cada valor de Y también se consideran haber sido tomados al azar de una población normal. Por lo tanto, cuando analizamos un coeficiente de correlación asumimos que existe una "distribución normal bivariada" de los datos. También resulta de interés aclarar que, en el caso de que la distribución fuese nonormal, los efectos adversos producidos por esto no serían compensados con un incremento en el tamaño de la muestra. Algunas de las "soluciones" estadísticas para resolver casos como el explicado podrían ser: a) antes de realizar las correlaciones observar si los datos en cada una de las variables se distribuyen de manera normal. Si estos se distribuyen de esa manera, el coeficiente de correlación tendrá una real validez. Si así no ocurriese, verificar nuevamente la normalidad de los datos después de la transformación de los mismos (esto es válido para las dos variables consideradas). Una transformación logarítmica (decimal o natural), la raíz cuadrada, una transformación angular, etc, generalmente son las más recomendadas, aunque el uso de cada una de ellas sea aconsejado para distintas situaciones; b) si las condiciones dadas en el punto (a) no ocurriesen, la manera más adecuada para el procesamiento de los datos seria la utilización de un método no paramétrico (por ej., el índice de correlación de Spearman, o el Coeficiente de correlación de Kendall, aunque el primero es más recomendable por su facilidad en el cálculo principalmente cuando el numero de datos es elevado [n= ≥ 30]).

Los test no paramétricos, en general, parten de la premisa que las observaciones bivariadas son mutuamente independientes, proviniendo cada una de la misma población continua (Potvin y Roff, 1993). Esta característica explica el porqué son tan utilizadas las correlaciones no paramétricas en biología, en contraposición al coeficiente de correlación producto-momento, que describe solamente la parte lineal de la relación entre dos variables. Al mismo tiempo, las técnicas estadísticas basadas en la clasificación u ordenación por rangos (por ejemplo, el método de Spearman o el de Kendall, entre otros) poseen además otras ventajas. En estos, las varianzas estimadas basadas en rangos, son menos sensibles a aquellos valores ubicados en los extremos de una distribución. Esta característica no ocurre cuando las estimamos a partir de los datos originales (Hettmansperger y McKean, 1978).

Un método alternativo a aquellos ordenados por rangos, es aquel denominado "Transformación por Rangos" (Rank Transformation [RT]). Este método fue propuesto por Conover e Iman (1981) como un puente entre los métodos paramétricos y no paramétricos. Básicamente este consiste en reemplazar los datos originales por sus rangos, para luego aplicar un test paramétrico (t test, F test, etc.). De este modo, resulta más probable que se satisfaga la asunción de los métodos paramétricos sobre la homogeneidad de las varianzas.

Literatura Citada

CONOVER, W. J. & R. L. IMAN. 1981. Rank transformation as a bridge between parametric and nonparametric statistics. *American Statistician* 35: 124-133.

HETTMANSPERGER, T. P. & J. W. MCKEAN. 1978. Statistical inference based on ranks. *Psychometrika* 43: 69-79.

LEVENE, H. 1960. Robust tests for equality of variance: 278-292. En I. Olkin; S. G. Ghurye; W. Hoeffding; W. G. Madow & H.B. Mann (eds). Contributions to Probability and Statistics. Stanford Univ. Press, California.

MARTIN, P. & P. BATESON. 1993. *Measuring Behavior: an introductory guide*. 2nd

edition. Carnbridge University Press. 222 pp.

POTVIN, C. & D. A. ROFF. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics?. *Ecology* 74 (6): 1617-1628.

ROHLF, F. J. & R. R. SOKAL. 1969. Statistical Tables. W.H. Freeman and Company. 252 pp.

SCHULTZ, B. 1983. On Levene's test and other statistics of variation. *Evolutionary Theory* 6: 197-203.

SNEDECOR, G. W. & W. G. COCHRAN. 1980. Statistical Methods. 7th edition. Ames, IA: Iowa State University Press. 507 pp.

SOKAL, R. R. & F. J. ROHLF. 1981. *Biometry*. 2nd edition. San Francisco: W.H. Freeman. 859 pp.

ZAR, J. H. 1984. *Biostatistical Analysis*. 2nd edition. New York: Prentice-Hall. 718 pp.